MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide

Vetle I. Torvik
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St.
Champaign, IL 61820
vtorvik@illinois.edu

ABSTRACT
Bibliographic records often contain author affiliations as free-form text strings. Ideally one would be able to automatically identify all affiliations referring to any particular country or city such as Saint Petersburg, Russia. That introduces several major linguistic challenges. For example, Saint Petersburg is ambiguous (it refers to multiple cities worldwide and can be part of a street address) and it has spelling variants (e.g., St. Petersburg, Sankt-Peterburg, and Leningrad, USSR). We have designed an algorithm that attempts to solve these types of problems. Key components of the algorithm include a set of 24k extracted city, state, and country names (and their variants plus geocodes) for candidate look-up, and a set of 1.1M extracted word n-grams, each pointing to a unique country (or a US state) for disambiguation. When applied to a collection of 12.7M affiliation strings listed in PubMed, ambiguity remained unresolved for only 0.1%. For the 4.2M mappings to the USA, 97.7% were complete (included a city), 1.8% included a state but not a city, and 0.4% did not include a state. A random sample of 300 manually inspected cases yielded six incompletes, none incorrect, and one unresolved ambiguity. The remaining 293 (97.7%) cases were unambiguously mapped to the correct cities, better than all of the existing tools tested: GoPubMed got 279 (93.0%) and GeoMaker got 274 (91.3%) while MediaMeter CLIFF and Google Maps did worse. In summary, we find that incorrect assignments and unresolved ambiguities are rare (< 1%). The incompleteness rate is about 2%, mostly due to a lack of information, e.g. the affiliation simply says "University of Illinois" which can refer to one of five different campuses. A search interface called MapAffil is available from http://abel.lis.illinois.edu/; the full PubMed affiliation dataset and batch processing is available upon request. The longitude and latitude of the geographical city-center is displayed when a city is identified. This not only helps improve geographic information retrieval but also enables global bibliometric studies of proximity, mobility, and other geo-linked data.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; I.5.4 [**Pattern Recognition**]: Applications – *text processing.*

General Terms
Algorithms.

Keywords
PubMed, MEDLINE, digital libraries, bibliographic databases, author affiliations, geographic indexing, place name ambiguity, geoparsing, geocoding, toponym extraction, toponym resolution.

1. INTRODUCTION
While information retrieval systems have become increasingly sophisticated in topic-based searching, other aspects of the bibliographic record have received much less attention. The author affiliation is one such aspect. For example, in MEDLINE, the US National Library of Medicine (NLM)'s premier bibliographic database covering biomedical-related papers published since ~1950, every paper is manually indexed with MeSH, their controlled vocabulary, and Entrez-PubMed (http://pubmed.gov) maps user queries into this vocabulary. First in 1988, the NLM started systematically indexing author affiliations, and only for the first-listed authors. As a result, it is easy to find papers on a topic like cancer with high precision and recall but it is nearly impossible to come up with a query to capture papers from, say, the United Kingdom – out of all the affiliations our algorithm mapped to the United Kingdom only 14% explicitly mention "United Kingdom" (another 10% mention England, Northern Ireland, Scotland, or Wales). Our motivation for geocoding affiliations in PubMed goes beyond basic information retrieval – it stems from efforts to disambiguate author names (Torvik and Smalheiser, 2009) and plans to carry out author-centered, bibliometric studies that include dimensions of geographic proximity and movement, and other data that can be linked to geographical locations.

The problem addressed in this paper is as follows: given a free-form text string representing an author affiliation, output the name of the corresponding city (or similar locality) and its physical location (the longitude and latitude of its center). If the city cannot be inferred, then output the country, and state (or equivalent subdivisions) when possible. For example, given `"McGill University Clinic, Royal Victoria Hospital, Montreal"`, then output `"Montreal, QC, Canada"` and its city-center coordinates. It should be noted that affiliation strings have been tagged as such in the XML distribution of MEDLINE/PubMed so extracting the affiliation string from a larger body of text is not an issue addressed here.

Why focus on the city and not on a more precise location such as the street address? Our goal is to assign geocodes at a uniform level across a broad spectrum of bibliographic records from across the world, some very old and with limited information. We have estimated that street addresses are present in only ~10% of PubMed records. The city (or a similar locality), we hypothesize, can be inferred from an affiliation string in the great majority of cases.

Geoparsing refers to the process of extracting toponyms (names of places or geographical entities) from text which are then fed into a geocoder to identify the corresponding physical location on the globe. Geoparsing and geocoding are active research areas, and a variety of related tools are available online. GoPubMed® (Doms & Schroeder, 2005; http://www.gopubmed.com) provides faceted searching of PubMed with a focus on topics but also has cities assigned to records, although it is not clear whether their data is made available in bulk or not. NEMO (Jonnalagadda et al., 2010) performs clustering in order to disambiguate institution names in PubMed affiliations, an effort that is complementary to ours. GeoMaker (Heilmann, 2009; http://icant.co.uk/geomaker;) is open-source and leverages Yahoo! PlaceMaker's extensive resources on places, organizations, and zip codes. Other tools are open-source but designed for different genres: Carmen (Dredze et al., 2013) is designed to geocode Twitter messages based on content and information about the users, while CLIFF (Bhargava and D'Ignazio, 2014; http://cliff.mediameter.org;) is designed to extract and geocode all mentions of people, places, and organizations from English natural language text. CLIFF uses a named entity extractor coupled with GeoNames (http://www.geonames.org) a large database of millions place names but we found that this can introduce unnecessary ambiguities and produce strange results: ”Abteilung fur Allergie und klinische Immunologie, Kinderklinik, Universitat La Sapienza, Roma” incorrectly mapped to ”Baden-Wurttemberg, Germany”, while “Victoria Hospital, London, Ont” incorrectly mapped to ”London, UK”. To be fair, GoPubMed got the same result in the latter case, and for the first case, GeoMaker returned nothing while Google Maps incorrectly returned a map of ”Erlangen, Germany”. These cases suggest that state-of-the-art tools are susceptible to systematic errors, rates of which we will estimate here, and compare to our own approach.

2. DATA AND METHODS

PubMed, which is the subject of this investigation, is a superset of MEDLINE – it covers older papers and out-of-scope journals and has records without MeSH but otherwise has metadata similar to MEDLINE, including affiliations. As mentioned, the NLM started systematically indexing affiliations of the first-listed authors in 1988. However, not all publishers provide affiliations in the records submitted to the NLM, and their indexing policy has changed over time (for a summary see the MEDLINE/PubMed Data Element Descriptions page; http://www.nlm.nih.gov/bsd/mms/medlineelements.html). As examples: starting in 1995, USA was added to the end of affiliations when deemed appropriate; starting in 1996, email addresses were appended, and in 1999, NLM stopped editing affiliations to "delete street information or redundant data" (NLM Tech Bull, 1999). In 2013, they stopped efforts to edit and quality control affiliations (NLM Tech Bull, 2013), and in 2014, moved the affiliation XML node from being linked to a paper to being linked to an author on a paper (NLM Tech Bull, 2014).

At the outset, we find that there is no typical affiliation string in PubMed: The majority are semi-structured (76% contain 3 or more commas, often used to separate department, institution, city, and state/country, in that order); many are non-English (~12% of university mentions are non-English like Universitat, Universite, Universidad, Uniwersytet); many are very short (4% have 40 or fewer characters, including punctuation); most are recent but some date all the way back to 1867; many common place names are ambiguous (Paris, London, Washington, New York, LA, Cambridge, and Boston all are), some more than others (e.g., Johnson, Union, and University are names of places); all affiliation strings are subject to errors due to the authors, copy-editing, character encoding, transliteration, and the indexing practices at the NLM.

Our approach is to take the affiliation at face value. That is, we do not use any external information attached to (or inferred from) the bibliographic record like the journal's country of publication, or other papers by the same author. However, this information could be used as a further step to help resolve remaining ambiguities, or infer a city when none is found. Although the final product is an entirely computational approach to mapping affiliation strings to a city, the design process necessitated significant manual effort. Several aspects of the algorithm, including the following two tasks, were refined after processing the entire collection of PubMed affiliations multiple times.

Task 1. Constructing a dictionary of city names, including known variant names, historical names, and misspelled variants, and their geocodes.

First a list of country names (and variants) and US states was constructed by studying the ending of all affiliations in the collection. Google Maps was used as a first pass on chunks extracted from affiliations that followed a certain structure that included the name of a country after the final comma, where the preceding two chunks, separated by commas, were submitted together with the country name as input to the Google Maps API. The two preceding chunks were used because many countries have a hierarchical structure much like the US: City, State, Country. As a result of this process, city names that never appeared in affiliations with this structure were not recorded during the first pass. As the algorithm and dictionary were iteratively refined, n-grams separated by commas in affiliations that were not assigned a city were collected and ranked by frequency, and then manually inspected in order to identify names of the most common cities missing from the dictionary. When Google Maps was unable to find the city, other resources were used on a case-by-case basis. Importing all the records of large-scale global resource of place names, like GeoNames, was considered but excluded in order to limit the overall ambiguity.

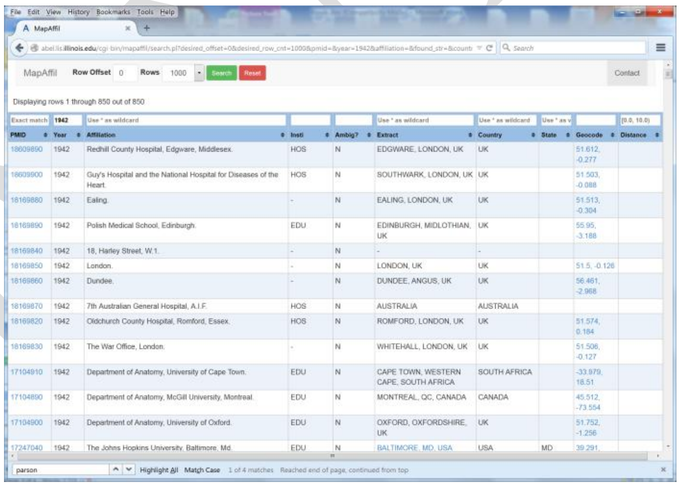
Task 2. Constructing a dictionary of word n-grams that (almost) uniquely point to a country (or US state).

All affiliation strings that were assigned to exactly one country were lowercased and all punctuation except space was removed. All 1-, 2-, 3-, and 4-grams that appeared on at least 3 different records were collected, and further filtered by restricting to n-grams that were 99% correlated with one specific country. For the USA, this process was repeated for its states and territories. This produced a total of 1.1M n-grams that almost exclusively point to a country, and when the country is the USA, can point to a US state or territory. For example, the 2-gram "iii friedrich" points to Germany. This list helps not only remove ambiguity in city names but also permits assigning an affiliation to a country when no place names is mentioned. Keep in mind that it is possible that a particular affiliation contains n-grams that point to multiple countries, particularly long unusual affiliations, but, as we shall see, it is rather rare that this phenomenon co-occurs with an otherwise unresolved ambiguity. Also, shorter affiliations are less likely to contain an n-gram from the dictionary, and as such are harder to disambiguate. It should also be noted that the n-gram dictionary is not the only manner in which the list of candidate places is refined, and ambiguity in place names is not the only phenomenon that creates a multiple candidate places.

```
21993610: Medicine and Pharmacology, Clinical Pharmacology and Hypertension, 1101 East Marshall Street, Sanger Hall, Room 8-062, Richmond, USA, dsica@mcvh-vcu.edu.
MapAffil: RICHMOND, VA, USA (77.433,37.541)
8939791: High Level Research1251 Mountain View DriveSmithfield, Utah 84335, USA.
MapAffil: SMITHFIELD, UT, USA (-111.825,41.832)
2725440: Department of Pharmacology, School of Pharmacy, University of Mississippi, University 38677
MapAffil: UNIVERSITY, MS, USA (-89.539,34.366)
9205386: Boston Education Centre, Pilgrim Hospital, Lincolnshire, USA.
MapAffil: BOSTON, LINCOLNSHIRE, UK (-0.004,52.976)
20101189: Department of Medicine, Montreal General Hospital and McGill University School of Medicine, Montreal, CA, USA.
MapAffil: MONTREAL, QC, CANADA (-73.554,45.512)
1628053: Health Centre, Thornaby, Cleveland.
MapAffil: THORNABY-ON-TEES, STOCKTON-ON-TEES, NORTH YORKSHIRE, UK (-1.298,54.538)
18446511: Center for Veterinary Medicine, The Food and Drug Administration, 7500 Standish Place, HFV-130, Rockville, Massachusetts 20855, USA.
MapAffil: ROCKVILLE, MD, USA (-77.151,39.082)
15694059: Coordinacion de Unidades de Medicina de Alta Especialidad, IMSS, Durango 289, 4 piso, Col. Roma, 06700 Mexico DF.
MapAffil: CUAUHTEMOC, CIUDAD DE MEXICO, DF, MEXICO (-99.144,19.443)
23393832: Iedico del Lavoro Competente, Tremestieri Etneo (CT), Italy
MapAffil: CATANIA, SICILIA, ITALY (15.088,37.503)
2265365: Vsezvazoveho vedeckeho centra lekarsko-biologickych problemov narkologie Ministerstva zdravotnictva ZSSR v Moskve.
MapAffil: MOSKVA, RUSSIA (37.618,55.756)
2799335: Rheumaklinik des Bethesda-Spitals Basel.
MapAffil: BASEL, SWITZERLAND (7.581,47.56)
```

Figure 1. A list of non-trivial affiliation strings with MapAffil output shown in red.

Assuming that two preceding dictionaries are in place, we can now describe the mapping algorithm. What follows is a brief outline because of space limitations but further details are available upon request. The first step involves pre-processing, chunking, and filtering the affiliation string, with the hopes that one or more of the chunks contain exact place names. A few of the highlights include converting all UTF-8 and html to ASCII, converting affiliations with all capital letters to first cap words, expanding some pairs of parentheses, introducing commas in strategic places into affiliations with no punctuation, collapsing chunks across commas when the resulting chunk leads to a valid place name, removing text that looks like a long narrative, extracting hand-coded patterns of country-specific zip codes, email addresses, urls, phone numbers, and street addresses. Once the pre-processing is finished, chunks of words that appear between commas are scanned for exact places names and placed on a high priority candidate list. A separate candidate list of lower priority is made up of place names that are a partial match within the chunks. These two candidate lists are then aligned with the countries and US states inferred from the word n-gram dictionary, zip code pattern, and email address in order to resolve part-of relations and prioritize the candidates. Candidates that appear further to the right in the affiliation are given higher weight, unless they are country names, as are the candidates on the exact match list compared to the partial match list. The final component of the overall algorithm is a short list of manually hard-coded rules that override some of the assignments made by this automatic process. These include cases of extreme ambiguity and ambiguities that are hard to resolve otherwise such as "`University, MS, USA`", and "`Ibaraki Prefecture, Japan`" vs. "`Ibaraki, Osaka, Japan`", and avoid mapping "`Harvard University`" or "`Harvard Medical School`" to "`Harvard, MA, USA`" unless it explicitly says so. Figure 1 provides a short list of non-trivial examples and their final successful assignments. Figure 2 shows the web-interface in use. Note the information sparsity in earlier records compared to more recent ones.



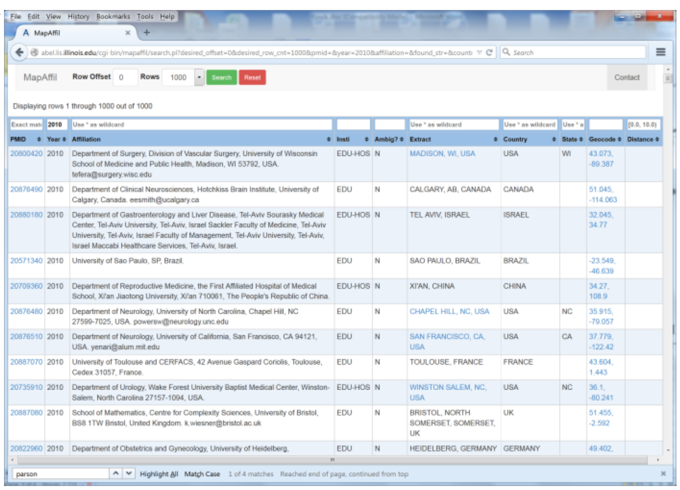


Figure 2. Screenshots of the MapAffil web-interface to PubMed records using publication year as input (top figure shows 1942; bottom figure shows 2010). All fields are searchable -- the affiliation field has been text-indexed using Sphinx for MySQL. Records include links to PubMed (via PMID), Google Maps (via geocodes for cities), and a summary of the 2010 US Census data (via FIPS code of the county that includes the geocode). Columns are included for institution type and note whether ambiguity was unresolved or not.

3. RESULTS

The algorithm was implemented using Perl because of extensive use of regular expressions. The implementation has not been optimized for speed but was fast enough to process 12.7 million affiliations in less than a week using a 32-core server. Table 1 shows a summary of the countries found in the collection of PubMed papers processed. Note that the bulk of the records start in 1988 (when the NLM started indexing affiliations in MEDLINE) but go back as far is 1867 partly because PubMedCentral is included in PubMed. The USA is by far the most frequent overall but is not as dominant in recent years.

Table 1. Worldwide distribution of 12.7M PubMed papers.

4163364 USA	3106 NEPAL	236 AZERBAIJAN
947014 JAPAN	2961 PERU	234 MOLDOVA
924305 UK	2944 INDONESIA	220 NICARAGUA
742280 GERMANY	2893 BOSNIA & HERZEGOVINA	218 BRUNEI
557106 CHINA	2693 TANZANIA	217 FIJI
515369 FRANCE	2599 SULTANATE OF OMAN	210 CENTRAL AFRICAN REPUBLIC
477050 ITALY	2531 SENEGAL	209 PARAGUAY
462732 CANADA	2466 UGANDA	206 LAOS
313557 SPAIN	2375 CAMEROON	188 MAURITIUS
299037 AUSTRALIA	2255 ZIMBABWE	174 GUINEA-BISSAU
275644 NETHERLANDS	2244 PHILIPPINES	148 NETHERLANDS ANTILLES
243401 INDIA	2173 GHANA	138 HONDURAS
202719 SWEDEN	2129 VIET NAM	136 GREENLAND
180918 BRAZIL	2002 JAMAICA	134 SIERRA LEONE
180437 KOREA	1907 ALGERIA	130 MONTENEGRO
168883 SWITZERLAND	1878 BELARUS	129 NAMIBIA
127352 TAIWAN	1685 COSTA RICA	129 MONGOLIA
126914 BELGIUM	1648 SUDAN	123 HAITI
126764 TURKEY	1613 IRAQ	122 DOMINICAN REPUBLIC
126307 POLAND	1543 QATAR	121 GUINEA
116470 ISRAEL	1457 REPUBLIC OF GEORGIA	100 AFGHANISTAN
112690 DENMARK	1418 COTE D'IVOIRE	96 BURUNDI
89686 FINLAND	1360 CYPRUS	93 EL SALVADOR
84958 AUSTRIA	1336 LUXEMBOURG	87 MAURITANIA
71705 NORWAY	1263 TRINIDAD & TOBAGO	85 KYRGYZSTAN
64848 GREECE	1244 MALAWI	82 LIECHTENSTEIN
62146 RUSSIA	1151 LATVIA	77 DJIBOUTI
52585 CZECH REPUBLIC	1066 MACEDONIA	74 SAINT KITTS & NEVIS
51329 MEXICO	1066 BURKINA FASO	65 CHAD
49781 NEW ZEALAND	995 ARMENIA	60 LESOTHO
49481 IRAN	945 PAPUA NEW GUINEA	60 BERMUDA
47475 HONG KONG	903 ZAMBIA	56 SWAZILAND
43693 HUNGARY	889 GAMBIA	55 SOMALIA
43353 ARGENTINA	885 PANAMA	54 ANGOLA
41013 SOUTH AFRICA	850 ECUADOR	52 ISLE OF MAN
39753 IRELAND	791 BAHRAIN	47 ERITREA
39653 SINGAPORE	788 MALTA	47 BHUTAN
39277 PORTUGAL	767 MADAGASCAR	46 SURINAME
30930 THAILAND	721 GABON	45 VANUATU
24804 EGYPT	711 SYRIA	45 FAEROE ISLANDS
23124 SAUDI ARABIA	710 LIBYA	31 ANDORRA
20216 CHILE	660 PALESTINE	30 SEYCHELLES
19504 NIGERIA	658 GUADELOUPE	30 SAMOA
18453 MALAYSIA	638 D.R. CONGO	26 SAN MARINO
17433 CROATIA	566 GUATEMALA	23 MALDIVES
16765 SERBIA	541 BENIN	21 EQUATORIAL GUINEA
16078 ROMANIA	531 MACAO	20 EAST TIMOR
15810 SLOVAKIA	529 MALI	20 ARUBA
15552 PAKISTAN	504 BOTSWANA	19 SAINT LUCIA
13081 TUNISIA	493 FRENCH GUIANA	17 COMOROS
11163 SLOVENIA	465 YEMEN	16 GIBRALTAR
9278 BULGARIA	456 MARTINIQUE	15 BELIZE
7735 COLOMBIA	445 TOGO	14 TONGA
7467 UKRAINE	427 UZBEKISTAN	13 TURKMENISTAN
7026 MOROCCO	422 MOZAMBIQUE	9 BRITISH VIRGIN ISLANDS
7001 VENEZUELA	421 CONGO	8 NORTH KOREA
6149 KENYA	387 KOSOVO	8 HOLY SEE
5754 LEBANON	386 BARBADOS	7 SAO TOME & PRINCIPE
5734 CUBA	366 CAMBODIA	5 TUVALU
5198 KUWAIT	358 MONACO	4 VATICAN CITY
4901 JORDAN	357 BOLIVIA	4 SAINT VINCENT & THE GRENADINES
4572 LITHUANIA	340 NEW CALEDONIA	4 CAPE VERDE
4521 BANGLADESH	305 NIGER	3 SAINT MARTIN
4054 ESTONIA	278 RWANDA	3 MONTSERRAT
3744 UNITED ARAB EMIRATES	278 KAZAKHSTAN	3 COOK ISLANDS
3528 ICELAND	276 GRENADA	2 WALLIS & FUTUNA
3367 ETHIOPIA	275 ALBANIA	1 TURKS & CAICOS ISLANDS
3285 URUGUAY	269 MYANMAR	1 SAINT PIERRE & MIQUELON
3195 SRI LANKA	268 FRENCH POLYNESIA	1 NIUE

Table 2 shows the results of head-to-head comparisons between MapAffil and four other tools: GoPubMed, GeoMaker, Google Maps, and CLIFF. These experiments were carried out using the respective web-based interfaces during a period of several days in May, 2015: http://www.gopubmed.com, http:// http://icant.co.uk/geomaker, http://maps.google.com, and http://cliff.mediameter.org; a link to GitHub suggested that CLIFF version 2.1.1 was running on the back-end. A strict definition of correct, unambiguous city was used. For example, inferring `London, UK` from "`Department of Agricultural Sciences, Imperial College London, Wye TN25 5AH, UK`" was judged incorrect even though the correct location `Wye, Ashford, Kent, UK` is near `London, UK`. However, inferring an alternative name for the correct city was judged correct, as was inferring a more precise location, such as a district or suburb within the correct city. Failure to resolve trivial part-of relations, as was often the case for CLIFF and GeoMaker, were judged correct instead of ambiguous. For example, it was judged correct when GeoMaker mapped "`Division of Cell Biology, Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands`" to both "`Amsterdam, North Holland`", `NL` and "`Netherlands`".

Table 2. Estimated performance rates based on a random sample of 300 affiliations. A smaller random subset of cases was deemed sufficient for estimating performance of Google Maps and CLIFF because their errors were not rare. *Note that GeoMaker and Google Maps had no ambiguous mappings by our design -- the top ranked result was taken for each query, otherwise the majority their results would be judged ambiguous.

	MapAffil	GoPubMed	GeoMaker	Google Maps	CLIFF
Correct Unambiguous City	293 **(97.7%)**	279 **(93.0%)**	274 **(91.3%)**	86 **(65.2%)**	77 **(58.3%)**
Incorrect	0	6	19	12	4
Ambiguous	1	0	0*	0*	5
None	1	2	5	33	10
State	4	12	2	0	9
Country	1	1	0	0	26
Total	300	300	300	132	132

GoPubMed represents an approach tailored specifically to PubMed affiliations -- each PubMed Identifier (PMID) was entered in their faceted interface and the mapped city looked-up in their "Locations" category. This does not explicitly give a longitude-latitude pair but rather a point on a small map and the name of the location which was used for these comparisons. After MapAffil, GoPubMed had the strongest performance: 93.7% of our test cases were correctly and unambiguously mapped to a city, compared by nearly 97.8% for MapAffil. The other tools had worse performance, which reflect generic efforts that have not been tailored to the specific genre analyzed here -- the author affiliations listed in PubMed.

Most of MapAffil's incomplete mappings were due to incomplete information available in the affiliation: "`Department of Emergency Medicine`." produced no output in all tools except Google Maps, which mapped it to Honolulu, HI, USA because of the present author's prior search history. Here are some other incomplete examples: "`Department of Laboratory Medicine, McMaster Medical Unit, Ontario, Canada.`", "`Department of Pediatrics, University of Kentucky, USA.`" Some of the cases that GoPubMed got wrong or incomplete include "`School of Pharmacy, Wingate University, Wingate, NC, USA.`" which it mapped to NC, USA. Furthermore, "Halso- och sjukvaardsnamndens forvaltning, Stockholms lans landsting." refers to Stockholm, Sweden but was mapped to Lens, France; "Japan Science and Technology Agency, Ishikawa, 923-1211, Japan." refers to Nomi City, Ishikawa Prefecture, Japan but was mapped to Ishikawa City, Okinawa Prefecture, Japan. Google Maps got both of these right, while MapAffil got the first one right and the second ambiguous (it identified both Ishikawa, Japan and Ishikawa, Okinawa, Japan), while CLIFF returned nothing for the first one and just Japan for the second one.

All geocoders were fed unedited affiliation strings. Google Maps and CLIFF could have performed better with some tweaking. For example, Google Maps tends to get overwhelmed and return "We could not find…." when given too much highly specific information such an email address and the name of a department within an institution. However, settings aside the 33 cases that returned "We could not find", still produces a high rate of incorrect mappings (12/(132-33) = 12.1%) because it appears to put more weight on names of institutions than names of places. CLIFF often removed names of organizations and people from the list of candidate places (e.g., Ann Arbor mapped to a person so was excluded as a city). With a little tweaking and pre-processing input given to both tools could help improve performance dramatically. GeoMaker uses information that is similar to that of Google Maps (names of institutions, places, and zip codes) except from a different source (Yahoo! PlaceMaker) and it refines the input/output.

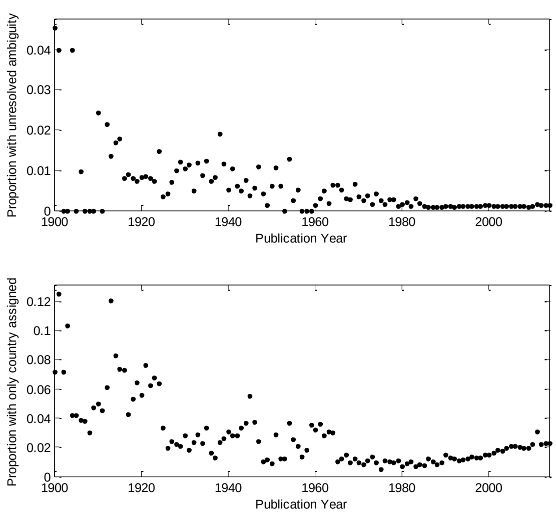


Figure 3. Unresolved ambiguity and incompleteness over time.

However, there was one case that CLIFF got complete and correct (mapped to Lake Worth, FL, USA) while few of the others did: "Kathleen D. Schaum, MS, is President and Founder of Kathleen D. Schaum&Associates, Inc, Lake Worth, Florida. Ms Schaum can be reached for questions and consultations by calling 561-964-2470 or through her e-mail address: kathleendschaum@bellsouth.net. Submit your questions for Payment Strategies by mail to Kathleen D. Schaum, MS, 6491 Rock Creek Dr, Lake Worth, FL 33467. Information regarding payment is provided as a courtesy to our readers, but does not guarantee that payment will be received. Providers are responsible for case-by-case documentation and justification of medical necessity." Google Maps timed out, GoPubMed returned As Sanamayn, Daraa, Syria, while MapAffil said USA because it filters out chunks of text that appears to be regular sentences.

When applied to a collection of 12.7 million affiliation strings listed in PubMed, ambiguity remained unresolved for only 0.1%. For the 4.2 million mappings to the USA, 97.7% were complete (included a city), 1.8% included a state but not a city, and 0.4% did not include a state. Figure 3 shows the rates of unresolved ambiguity and incompleteness over time. Ambiguity has been very low since ~1980 but we see significant ambiguity in earlier papers. This is a reflection of how affiliations are written in earlier days. Figure 2 shows that affiliations from the 1940s are very short, sometimes even just listing the name of a city, compared to the longer ones of today that include departments, institutions, street addresses, cities, states, countries, zip codes, emails, and so on. We also observe that the incompleteness rate has been slightly but steadily increasing over time since 1980. This probably reflects an increasingly diverse set of affiliations. We also found about 40k affiliations that only listed an email address, and email addresses in affiliations have generally been on the rise.

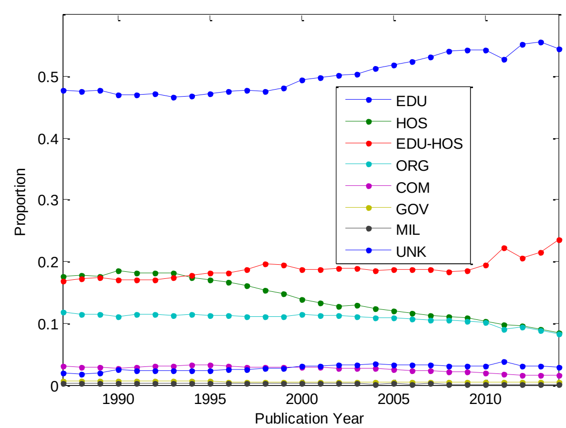


Figure 4. Affiliation types over time.

Affiliation types where captured using simple regular expressions into 8 different categories: EDUcational, HOSpital, EDUcationa-HOSpital, ORGanization, COMmercial, GOVernment, MILitary, UNKown. First the affiliation was matched against EDU or HOS, or both. If neither matched, then one other category was matched if possible. ORG represent a generic research organization, and includes national institutes/laboratories/centers, associations, etc. GOV includes institutions like local health departments but not national institutes, hospitals, or educational institutions. Figure 3 shows the prevalence of the different kinds of institutions over time in the dataset. The two dominant categories are educational institutions and hospitals. We have performed preliminary experiments on large collections of principal investigators and their affiliations listed in NIH and NSF grants, as well as inventors' addresses on USPTO patents. NIH and NSF are also dominated by education (and hospitals for NIH). The patent genre is quite different. Inventors often do not have an institutional affiliation, and their home addresses are listed, and the assignees are most often commercial entities. This makes the set of locations much more diverse. Even so, MapAffil presently covers greater than 90% of these records. We expect some of the more generic tools tested in our experiments to have higher coverage for USPTO inventor addresses but have not tested this yet.

4. DISCUSSION

As mentioned earlier the current algorithm is the result of several iterations of refinement. At this point the accuracy of the algorithm has plateaued, in the sense that major new components are necessary to significantly improve performance. Adding a thousand new (rare) cities to the locations dictionary would have little effect on overall performance. We find that incorrect assignments and unresolved ambiguities are rare (< 1%). The incompleteness rate is about 2%, mostly due to a lack of information. In order to improve completeness in these cases, one could include information external to the affiliation field such as other papers by the same author or constructing a list of institutions that can be unambiguously mapped to one location. This information can be used both as a further step to help remove ambiguity or infer a city when only country is given.

Nevertheless, the current performance is much greater than other tools and should enable new types of global bibliometric studies on geographical proximity and geo-linked data. As examples, we are presently studying the impact of local demographics on the diversity of co-authorships and topics in biomedical science, and building models of collaborative behavior where geographical proximity is one of several important explanatory variables.

5. ACKNOWLEDGMENTS

Research reported in this publication was supported in part by the National Institute on Aging of the US NIH grant P01AG039347 and the Directorate for Education & Human Resources of the US NSF award 1348742. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF.

REFERENCES

[1] Jonnalagadda SR, Topham P. 2010. NEMO: Extraction and normalization of organization names from PubMed affiliations. *J Biomed Discov Collab. 2010 Oct 4;5:50-75*. DOI= http://dx.doi.org/10.5210/disco.v5i0.3047

[2] French JC, Powell AL, Schulman E. 2000. Using clustering strategies for creating authority files. *J. Am. Soc. Inf. Sci.*, 51: 774–786. DOI=http://dx.doi.org/ 10.1002/(SICI)1097-4571(2000)51:8<774::AID-ASI90>3.0.CO;2-P

[3] Torvik VI, Smalheiser NR. 2009. Author name disambiguation in MEDLINE. ACM TKDD 3(3): 11. DOI=http://dx.doi.org/10.1145/1552303.155230

[4] Dredze M, Paul MJ, Bergsma S, Tran H. 2013. Carmer: A Twitter geolocation system with applications to public health. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI, Bellevue, WA.

[5] Doms A, Schroeder M. 2005. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Research 33 (Web Server issue): W783-W786.

[6] GoPubMed by Transinsight GmbH: http://www.gopubmed.com; Accessed May, 2015.

[7] Zhang W, Gelernter J. 2014. Geocoding location expressions in Twitter messages: A preference learning method. Journal of Spatial Information Science. 9: 37-70.

[8] Leidner JL. 2007. Toponym resolution in text. PhD thesis, University of Edinburgh, UK.

[9] Heilmann C. 2009. GeoMaker. http://icant.co.uk/geomaker; Accessed May, 2015.

[10] Bhargava R, D'Ignazio C. 2014. CLIFF Mediameter. MIT Center for Civic Media. http://cliff.mediameter.org/; Accessed May, 2015.

[11] GeoNames. http://www.geonames.org/; Accessed May, 2015.

[12] NLM Tech Bull. 1999 Nov-Dec;(311). https://www.nlm.nih.gov/pubs/techbull/nd99/nd99_changes.html; Accessed May, 2015.

[13] NLM Tech Bull. 2013 Sep-Oct;(394):b4. http://www.nlm.nih.gov/pubs/techbull/so13/brief/so13_author_affiliations.html; Accessed May, 2015.

[14] NLM Tech Bull. 2014 Nov-Dec;(401):e5. http://www.nlm.nih.gov/pubs/techbull/nd14/nd14_medline_data_changes_2015.html; Accessed May, 2015.

[15] MEDLINE/PubMed Data Element (Field) Descriptions. http://www.nlm.nih.gov/bsd/mms/medlineelements.html; Accessed May, 2015.